Biterm Topic Model

Nguyễn Bá Cương

School of Information and Communication Technology Hanoi University of Science and Technology

Data Science Lab , 2017

Nội dung

Nội dung

Gới thiệu về short texts

Những văn bản ngắn rất là phổ biến trên các trang web

Ứng dụng của short texts

Hiểu được chủ đề của các văn bản ngắn rất là quan trong trong nhiều lĩnh vực

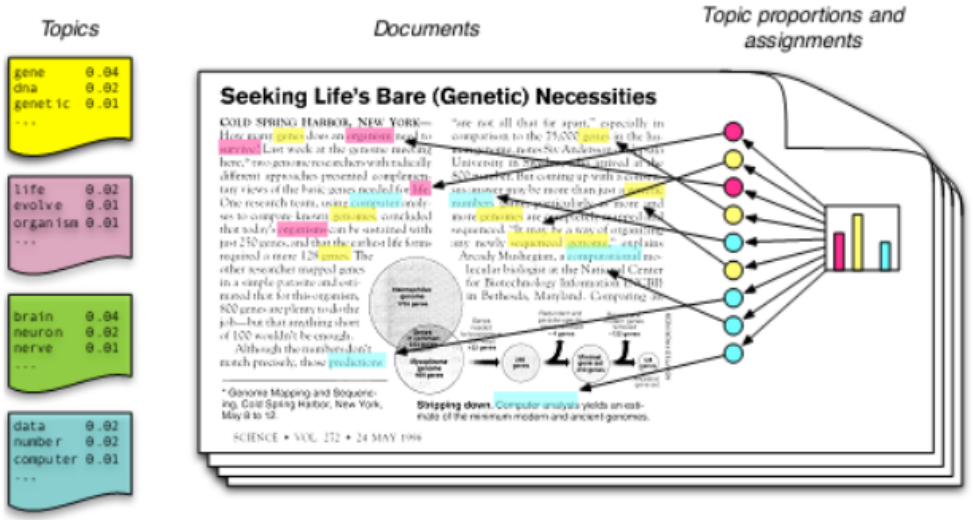
- Mô tả đặc điểm nội dung (content characterizing)
- Gợi ý nội dung (content recommendation)
- Sở thích người dùng (user interest profiling)
- Phát hiện các chủ đề nổi lên (emerging topic detecting)
- Phân tích ngữ nghĩa (semantic analysic)
- ...

Những vần đề của short texts

Những khó khăn và thách thức

- Không đủ ngữ cảnh để xác định ý nghĩa của câu
- Các đặc điểm short texts
  - Độ dài văn bản rất ngắn
  - Số lượng dữ liệu của văn bản ngắn là rất lớn và tăng nhanh
  - Các chủ để nó phản ánh đến các xu hướng xã hội
- Việc giới hạn độ dài của văn bản trong short texts làm cho chúng rất khó để phân tích với các mô hình xác suất truyền thống

Mô hình chủ đề



From Blei

(a) Mô hình chủ đề

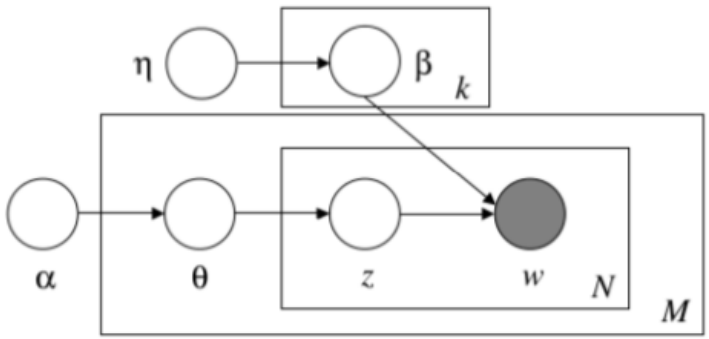
Mô hình sinh của các văn bản với chủ đề ẩn

- Một chủ đề $\sim$ một phân phối xác suất trong tập từ vững
- Một văn bản $\sim$ một tập trộn của các chủ đề ẩn
- Một từ $\sim$ một điểm trong một chủ đề

Mô hình LDA

Mô hình sinh của mô hình LDA với K chủ đề.

- Sinh các phân phối từ theo chủ đề
  1. Với một chủ đề i trong trong K
  a). Lấy mẫu $\beta_i \sim Dir(\eta)$
- Sinh các từ của một văn bản
  1. Chọn phân phối trộn các chủ đề của văn bản $\theta \sim Dir(\alpha)$
  2. Ứng với một từ w_n trong văn bản
  a). Chọn một chủ đề $z_n \sim Mul(\theta)$
  b). Chọn ra một từ $w_n \sim Mul(\beta_{z_n})$



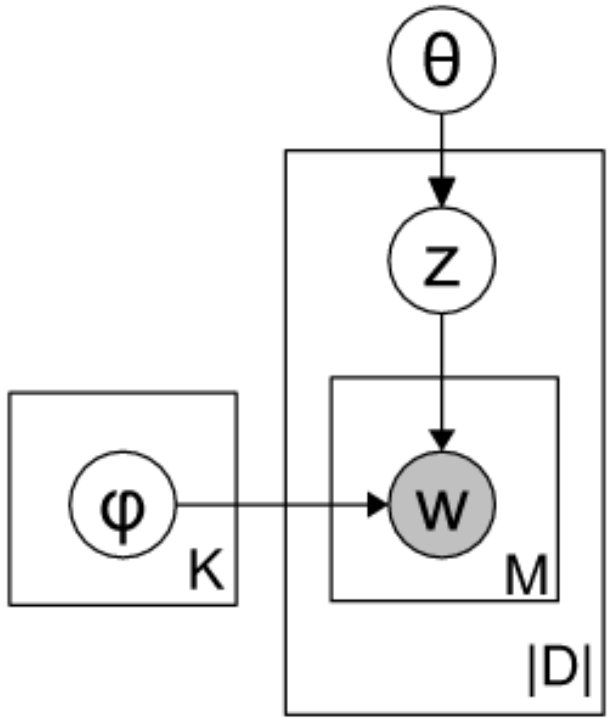
(b) Mô hình LDA

LDA cho short texts

- Vấn đề:
  - Độ dài văn bản ngắn, các từ hầu như chỉ xuất hiện một lần
  - Không đủ ngữ cảnh để xác định nội dung
- Một số giải pháp:
  - Khái thác kiến thức từ bên ngoài để làm giàu sự biểu diễn cho short texts
  - Kết hợp một số văn bản ngắn thành một văn bản dài dựa trên một số thông tin. Như tập hợp các bài đăng trên Twitter bởi cùng người người dùng...

Mô hình mixture of unigrams

- Giả định rằng tất cả các từ thuộc một văn bản đều thuộc cùng một chủ đề
  => Vấn đề: Mặc dù là short text nhưng mỗi văn bản vẫn có thể có nhiều hơn 1 chủ đề



(c) Mô hình Mixture of unigrams

Nội dung

Ý tưởng

- Chủ đề cơ bản là một nhóm các từ tương quan với nhau và những từ tương quan với nhau này được phát hiện bởi sự đồng xuất hiện trong một văn bản.
  $\rightarrow$ Tại sao không tồn tại mô hình đồng thời xuất hiện các từ để học từng chủ đề
- Mô hình chủ đề cho short texts chịu nhiều ảnh hưởng của dữ liệu thưa
  $\rightarrow$ Tại sao không sử dụng toàn bộ nguồn dữ liệu để học ra các chủ đề

Xây dựng biterm và dữ liệu học

- Biterm là một cặp từ không theo thứ tự cùng xuất hiện trong một văn bản

  Ví dụ như một văn bản gồn có 3 từ w_1, w_2, w_3 thì biterm là

  $$(w_1, w_2, w_3) \rightarrow \{(w_1, w_2), (w_2, w_3), (w_1, w_3)\}$$

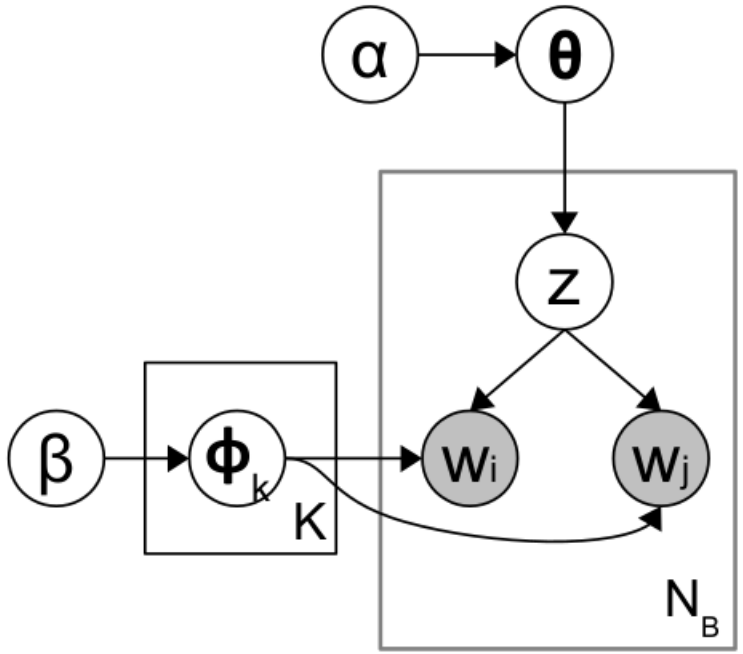
- Dữ liệu học bao gồm tất cả các biterm được sinh ra từ bộ dữ liệu ban đầu

Mô hình Biterm Topics Model (BTM)

Mô hình sinh của BTM

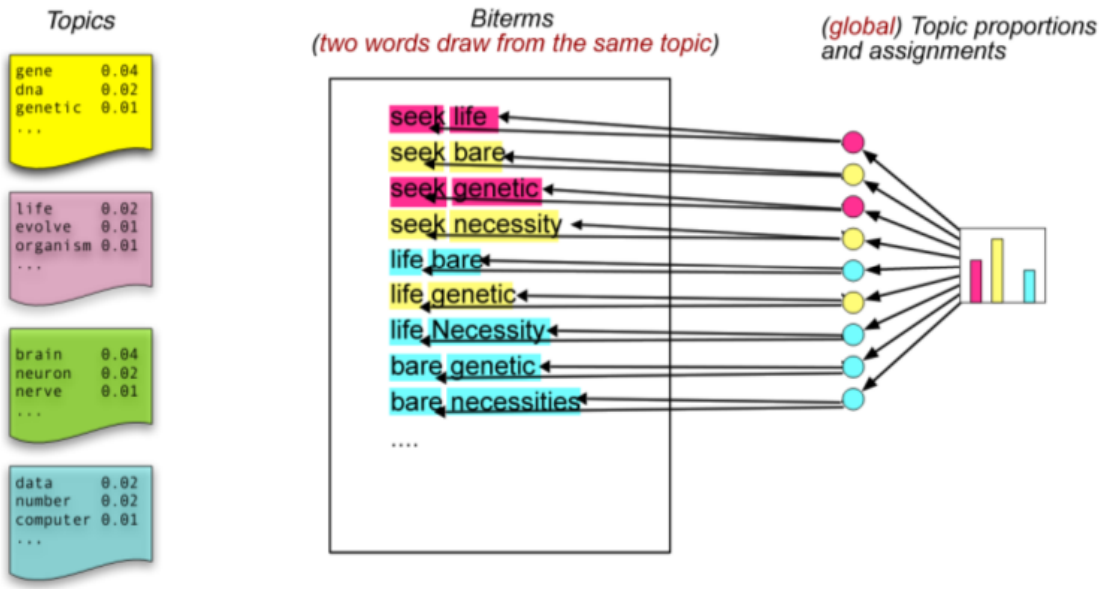
- Vời từng chủ đề z
  a). Lấy mẫu $\phi_z \sim Dir(\beta)$
- Chọn phân phối chủ đề $\theta \sim Dir(\alpha)$ cho toàn bộ tập dữ liệu
- Với từng biterm b
  a). Chọn một phân phối chủ đề $z \sim Mutil(\theta)$
  b). Chọn phân phối các từ trong biterm

$$w_1, w_2 \sim Mutil(\phi_z)$$



(d) Mô hình BTM

Mô hình BTM



Mô hình sinh của biterms với chủ đề ẩn

- Chủ đề $\sim$ phân phối qua các từ
- Tập dữ liệu $\sim$ tập trộn các chủ đề
- Một biterm $\sim$ hai từ xác định biểu diễn cùng trong một chủ đề

Suy diễn chủ đề cho từng văn bản

- Giả định
  Tỉ lệ chủ đề của một văn bản tương đương với kì vọng tỉ lệ chủ đề của các biterm trong văn bản đó.

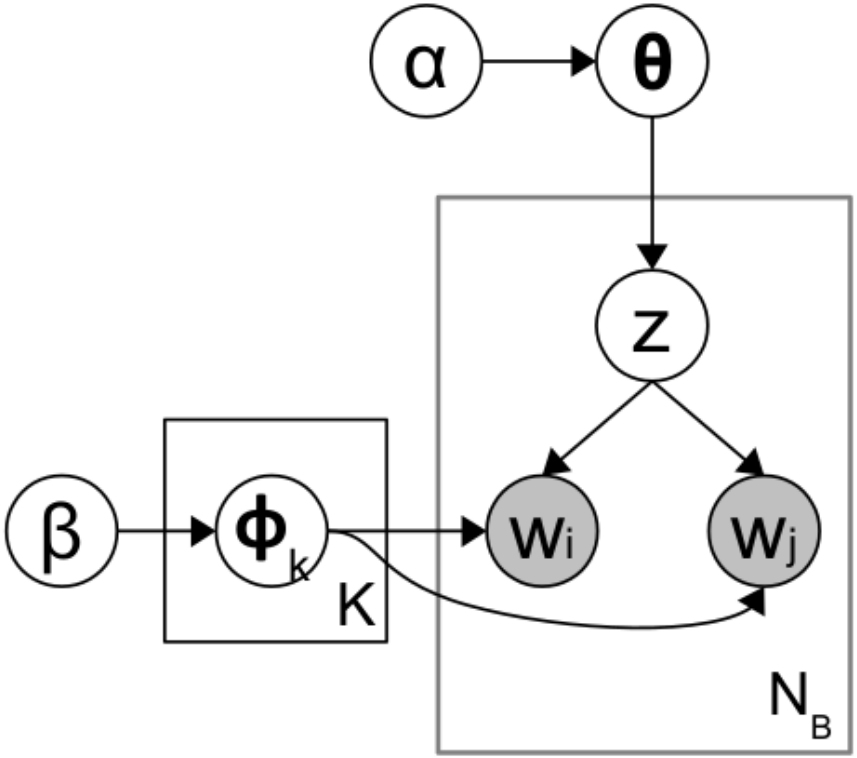
$$P(z|d) = \sum_b P(z|b)P(b|d)$$

- Trong đó

$$P(z|b) = \frac{P(z)P(w_|z)P(w_j|z)}{\sum_z P(z)P(w_i|z)P(w_j|z)}$$

$$P(b|d) = \frac{n_d(b)}{\sum_b n_d(b)}$$

Maximum A Posteriori



- α, β là hyperparameter. Ta xác định giá trị của nó bằng cross-validation
- Z là biến ẩn, chúng ta không quan sát được
- Tập các biterm B là dữ liệu duy nhất chúng ta quan sát được
- Φ, Θ là biến chúng ta cần tìm

Sử dụng ước lượng MAP:

$$\begin{aligned} \Phi, \Theta &= \text{argmax } \log P(\Phi, \Theta | \alpha, \beta, B) \\ &= \text{argmax } \log P(B | \Phi, \Theta) P(\Phi, \Theta | \alpha, \beta) \end{aligned} \quad (1)$$

Phương pháp Gibbs Sampling

- Ý tưởng: Sinh ra một mẫu phân bố hậu nghiệm bằng cách duyệt qua các giá trị từ tập dữ liêu ban đầu.
- Ví dụ:
  Cho các biến ngẫu nhiên X_1, X_2, X_3
  Khởi tạo: $x_1^{(0)}, x_2^{(0)}, x_3^{(0)}$
  Tại bước lặp thứ i

$$x_1^{(i)} = p(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)})$$
$$x_2^{(i)} = p(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)})$$
$$x_3^{(i)} = p(X_3 = x_3 | X_1 = x_1^{(i)}, X_2 = x_2^{(i)})$$

  Qúa trình này sẽ dừng đến khi hội tụ

Suy diễn tham số

Lấy mẫu chủ đề cho từng biterm trong mini-batch t

$$P(z_i = k|z_{-i}^{(t)}, B^{(t)}, \alpha^{(t)}, \{\beta_k^{(t)}\}_{k=1}^K) \propto$$
$$(n_{-i,k}^{(t)}, \alpha_k^{(t)}) \frac{(n_{-i,w_i|k}^{(t)} + \beta_{k,w_i}^{(t)})(n_{-i,w_j|k}^{(t)} + \beta_{k,w_j}^{(t)})}{[\sum_{w=1}^{W}(n_{-i,w|k}^{(t)} + \beta_{k,w}^{(t)})]^2} \quad (2)$$

Cập nhật lại các tham số

$$\alpha_k^{(t+1)} = \alpha_k^{(t)} + \lambda n_k^{(t)} \quad (3)$$

$$\beta_{k,w}^{(t+1)} = \beta_{k,w}^{(t)} + \lambda n_{w|k}^{(t)} \quad (4)$$

$$\phi_{k,w}^{(t)} = \frac{n_{w|k}^{(t)} + \beta^{(t)}}{n_{.|k}^{(t)} + W\beta^{(t)}} \quad (5)$$

$$\theta_k^{(t)} = \frac{n_k^{(t)} + \alpha^{(t)}}{N_\beta^{(t)} + K\alpha^{(t)}} \quad (6)$$

Algorithm 1 Thuật toán Gibbs sampling dạng online cho mô hình BTM

Input: K, α, β, biterm sets $B^{(1)}, ..., B^{(T)}$
Output: $\{\phi^{(t)}, \theta^{(t)}\}_{t=1}^{T}$

Set $\alpha^{(1)} = (\alpha, ..., \alpha)$ and $\{\beta_k^{(1)} = (\beta, ..., \beta)\}_{k=1}^{K}$
for $t = 1$ to T do **do**
    Randomly initialize the topic asignments for all the biterms
    for *iter* $= 1$ to N_{iter} do **do**
        for each biterm $b_i = (w_{i,1}, w_{i,2}) \ni B^{(t)}$ **do**
            Drawn topic k from Eq.(2)
            Update $n_k^{(t)}, n_{w_{i,1}|k}^{(t)}$ and $n_{w_{i,2}|k}^{(t)}$
        end for
        $\alpha^{(t+1)}$ and $\{\beta_k^{(t+1)}\}_{k=1}^{K}$ by Eq.(3)and Eq.(4)
    end for
    Compute $\phi^{(t)}$ by Eq.(5) and $\theta^{(t)}$ by Eq.(6)
end for

Vấn đề và phương pháp học mới

- Dữ liệu biterm sinh ra quá lớn
  $=>$ Việc sử dụng phương pháp Gibbs sampling không hiệu quả và tốn nhiều thời gian
- Phương pháp học mới
  - Phương pháp học VB
  - Phương pháp học ngẫu nhiên trên bộ dữ liệu

Tư tưởng của Expectation - Maximization

- Ước lượng MAP (1) trong trường hợp này rất khó khăn và không có công thức cụ thể
  $=>$ ta xây dựng một lower-bound cho hàm mục tiêu này và tối ưu hóa trên lower-bound này
- Tư tưởng thuật toán.
  Giả sử mô hình có tham số θ, biến quan sát được X và tập các biến ẩn Z $=>$ Sử dụng ước lượng MAP ta có:

$$\begin{aligned}\theta_{MAP} &= \text{argmax} \prod_{i}^{N} P(x_i|\theta)P(\theta) \\ &= \text{argmax} \sum_{i}^{N} \log P(x_i|\theta) + \log P(\theta) \\ &= \text{argmax} \sum_{i}^{N} \log \sum_{z_i} P(x_i z_i|\theta) + \log P(\theta) \qquad (7)\end{aligned}$$

Tư tưởng thuật toán EM

Với mỗi giá trị i, gọi $Q(z_i)$ là một phân phối trên biến ẩn z_i, áp dụng bất đẳng thức Jensen ta có:

$$\sum_i^N \log \sum_{z_i} P(x_i z_i|\theta) + \log P(\theta)$$

$$= \sum_i^N \log \sum_{z_i} Q(z_i) \frac{P(x_i z_i|\theta)}{Q(z_i)} + \log P(\theta)$$

$$\geq \sum_i^N \sum_{z_i} Q(z_i) \log \frac{P(x_i z_i|\theta)}{Q(z_i)} + \log P(\theta) = \mathcal{L}(Q, \theta)$$

Dấu bằng khi:

$$Q(z_i) \propto P(z_i|x_i, \theta)$$

Tư tưởng thuật toán EM

Hàm lower-bound:

$$\mathcal{L}(Q,\theta) = \sum_{i}^{N}\sum_{z_i} P(z_i|x_i,\theta^{old}) \log P(x_i z_i|\theta)$$
$$- \sum_{i}^{N}\sum_{z_i} P(z_i|x_i,\theta^{old}) \log P(z_i|x_i,\theta^{old}) + \log P(\theta)$$

- Bước E

$$\text{Xác định } \mathcal{Q}(\theta) = \sum_{i}^{N}\sum_{z_i} P(z_i|x_i,\theta^{old}) \log P(x_i z_i|\theta) + \log P(\theta)$$

- Bước M

$$\text{Tính } \theta^{new} = \text{argmax}_\theta \mathcal{Q}(\theta)$$

Áp dụng thuật toán cho mô hình BTM

Bước E ta tính giá trị

$$\begin{aligned} t_{n,k} &= P(z_n = k|b_n, \theta, \Phi) \\ &= P(z_n = k|w_{n,1}, w_{n,2}, \theta, \Phi) \\ &= \frac{P(w_{n,1}, w_{n,2}|z_n = k, \theta, \Phi)P(z_n = k|\theta, \Phi)}{\sum_k P(w_{n,1}, w_{n,2}|z_n = k, \theta, \Phi)P(z_n = k|\theta, \Phi)} \\ &= \frac{\phi_{k,w_{n,1}}\phi_{k,w_{n,2}}\theta_k}{\sum_k \phi_{k,w_{n,1}}\phi_{k,w_{n,2}}\theta_k} \end{aligned}$$

Bước M ta cập nhật giá trị

$$\theta_k = \frac{\sum_n t_{n,k} + \alpha}{\sum_{k'} \left(\sum_n t_{n,k'} + \alpha\right)}$$

$$\phi_{k,w} = \frac{\sum_n t_{n,k}\ c(b_n, w) + \beta}{W\beta + \sum_n 2t_{n,k}}$$

Algorithm 2 Thuật toán EM dạng online cho mô hình BTM

Define $\gamma_i = (i+2)^{-p}$

Input: topic number K, α, β, biterm sets $B^{(1)}, ..., B^{(T)}$

Output: ϕ, θ

1: Randomly initialize $\phi, \theta, S^0_{\theta_k} = 0, S^0_{\phi_{k,w}} = 0, i = 0$
2: **for** $i = 1$ to ∞ **do**
3: **for each** biterm $b_j = (w_{i,1}, w_{i,2}) \ni B^{(i)}$ **do**
4: $t_{j,k} \propto \phi_{k,w_{j,1}} \phi_{k,w_{j,2}} \theta_k$
5: **end for**
6: $S^i_{\phi_{k,w}} = \sum_j t_{i,k} c(b_j, w)$; $S^i_{\theta_k} = \sum_j t_{j,k}$
7: $S^i_{\phi_{k,w}} = (1-\gamma_i) S^{(i-1)}_{\phi_{k,w}} + \gamma_i S^i_{\phi_{k,w}}$; $S^i_{\theta_k} = (1-\gamma_i) S^{(i-1)}_{\theta_k} + \gamma_i S^i_{\theta_k}$
#Update
8: $\theta_k \propto S^i_{\theta_k} + \alpha$; $\phi_{k,w} \propto S^i_{\phi_{k,w}} + \beta$
9: **end for**

Phương pháp học ngẫu nhiên

Những vấn đề về:

- Dữ liệu biterm sinh ra quá lớn.
- Vấn đề về dữ liệu nhiễu.

=> Giải pháp đưa ra:
Học một cách ngẫu nhiên một tập dữ liệu trên tập dữ liệu gốc như sau:
Với mỗi minibatch chọn ngâu nhiên một lương biterm bằng cách sử dụng phân phối nhị thức với một xác suất p.

Algorithm 3 Thuật toán Online VB cho mô hình BTM

Define $\gamma_i = (i+2)^{-p}$

Input: topic number K, α, β, biterm sets $B^{(1)}, ..., B^{(T)}$

Output: ϕ, θ

Randomly initialize $\phi, \theta, S^0_{\theta_k} = 0, S^0_{\phi_{k,w}} = 0, i = 0$

for $i = 1$ to ∞ **do**

  Generating minibatch $B^{`(i)}$ form $B^{(i)}$

  for each biterm $b_j = (w_{i,1}, w_{i,2}) \ni B^{`(i)}$ **do**

    $t_{j,k} \propto \phi_{k,w_{j,1}} \phi_{k,w_{j,2}} \theta_k$

  end for

  $S^i_{\phi_{k,w}} = \sum_j t_{i,k} c(b_j, w)$; $S^i_{\theta_k} = \sum_j t_{j,k}$

  $S^i_{\phi_{k,w}} = (1-\gamma_i) S^{(i-1)}_{\phi_{k,w}} + \gamma_i S^i_{\phi_{k,w}}$; $S^i_{\theta_k} = (1-\gamma_i) S^{(i-1)}_{\theta_k} + \gamma_i S^i_{\theta_k}$

  #Update

  $\theta_k \propto S^i_{\theta_k} + \alpha$; $\phi_{k,w} \propto S^i_{\phi_{k,w}} + \beta$

end for

Algorithm 4 Online BTM Algorithm

Input: K, α, β, biterm sets $B^{(1)}, ..., B^{(T)}$
Output: $\{\phi^{(t)}, \theta^{(t)}\}_{t=1}^{T}$

Set $\alpha^{(1)} = (\alpha, ..., \alpha)$ and $\{\beta_k^{(1)} = (\beta, ..., \beta)\}_{k=1}^{K}$
for $t = 1$ to T do **do**
  Randomly initialize the topic asignments for all the biterms
  for *iter* $= 1$ to N_{iter} do **do**
    Generating minibatch $B^{`(t)}$ form $B^{(t)}$
    for each biterm $b_i = (w_{i,1}, w_{i,2}) \ni B^{`(t)}$ **do**
      Drawn topic k from Eq.(1)
      Update $n_k^{(t)}$, $n_{w_{i,1}|k}^{(t)}$ and $n_{w_{i,2}|k}^{(t)}$
    end for
    $\alpha^{(t+1)}$ and $\{\beta_k^{(t+1)}\}_{k=1}^{K}$ by Eq.(2)and Eq.(3)
  end for
  Compute $\phi^{(t)}$ by Eq.(4) and $\theta^{(t)}$ by Eq.(5)
end for

Nhận xét

- Việc học ngẫu nhiên một lượng biterm trên tập dữ liệu gốc
  $=>$ Cải thiện thời gian học
- Khắc phục được vấn đề nhiễu dữ liệu. Có những từ có thể không liên quan gì đến nhau có thể được loại bỏ trong quá trình chọn ngẫu nhiên.

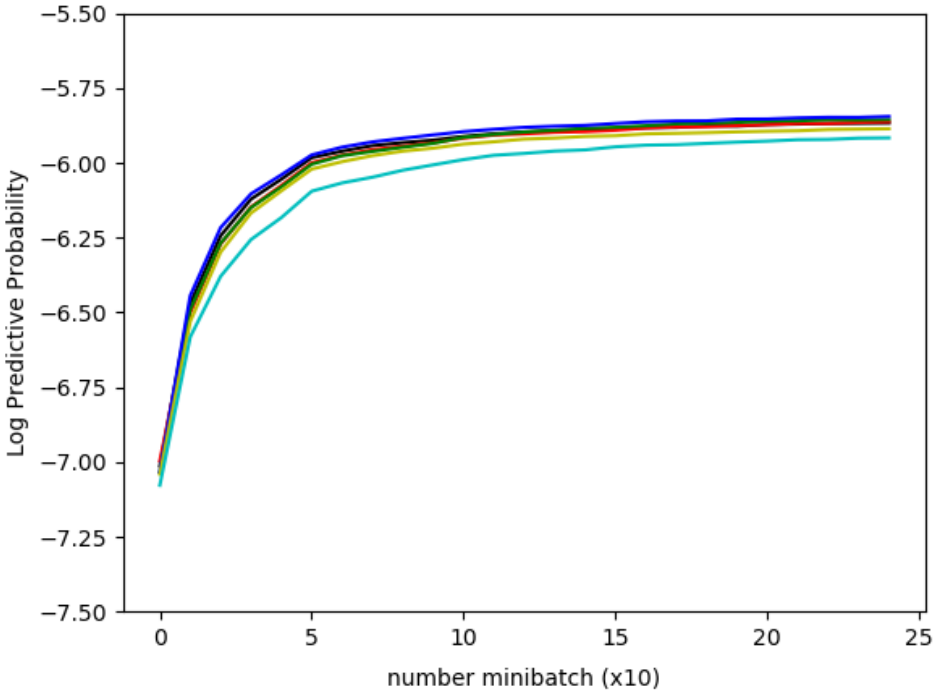
Nội dung

Bộ dữ liệu thử nghiệm

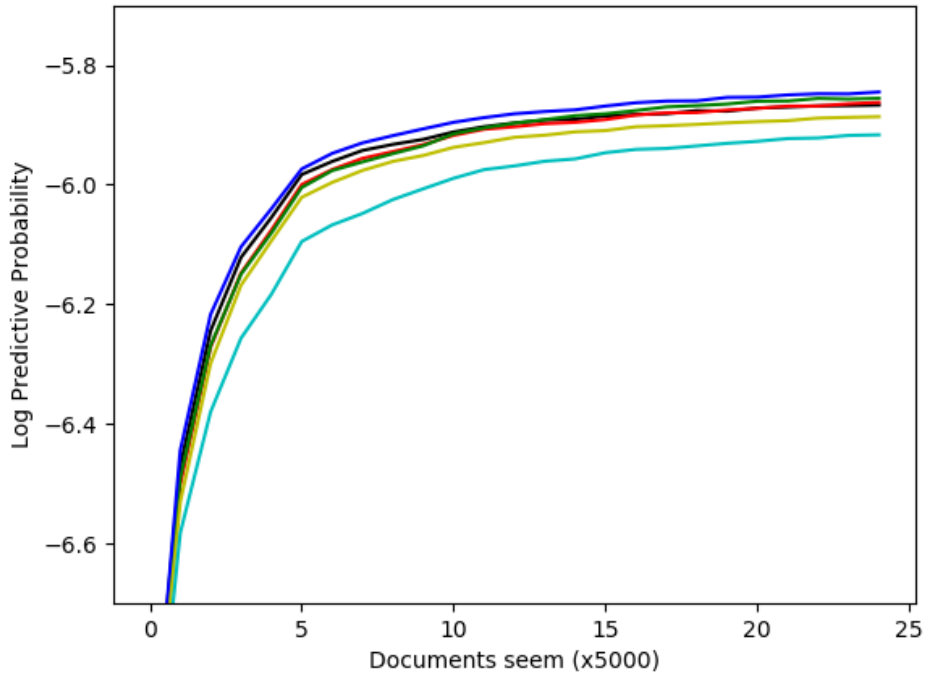
	Corpus size	Average length per doc	V
Yahoo Questions	537,770	4.73	24,420
Tweets	1,485,068	10.14	89,474
Nytimes Titles	1,684,127	5.15	55,488

Bảng : Bảng mô tả dữ liệu thử nghiệm

Tập dữ liệu Tweets, K =100, độ đo perplexity

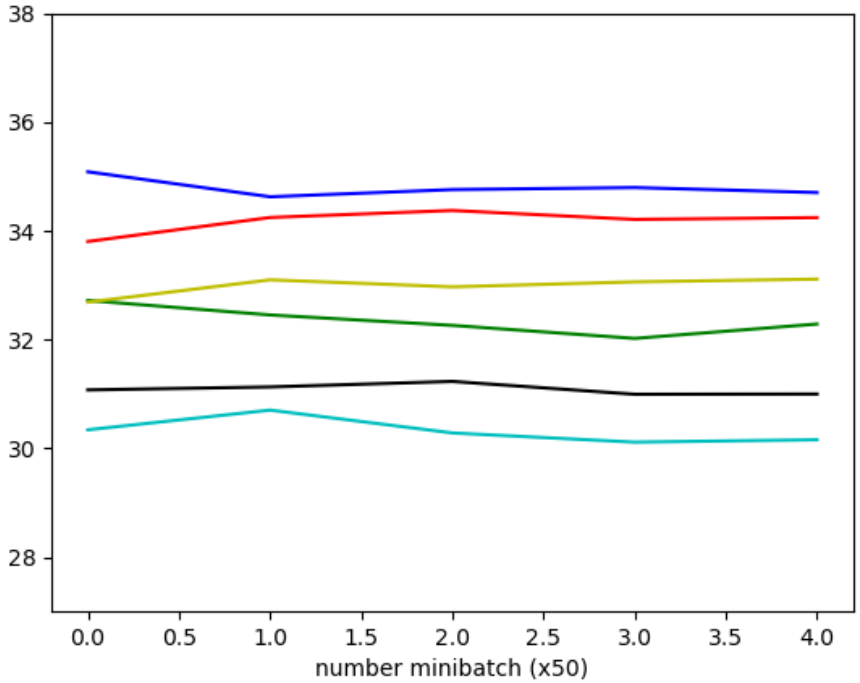


(e) Online Gibbs sampling

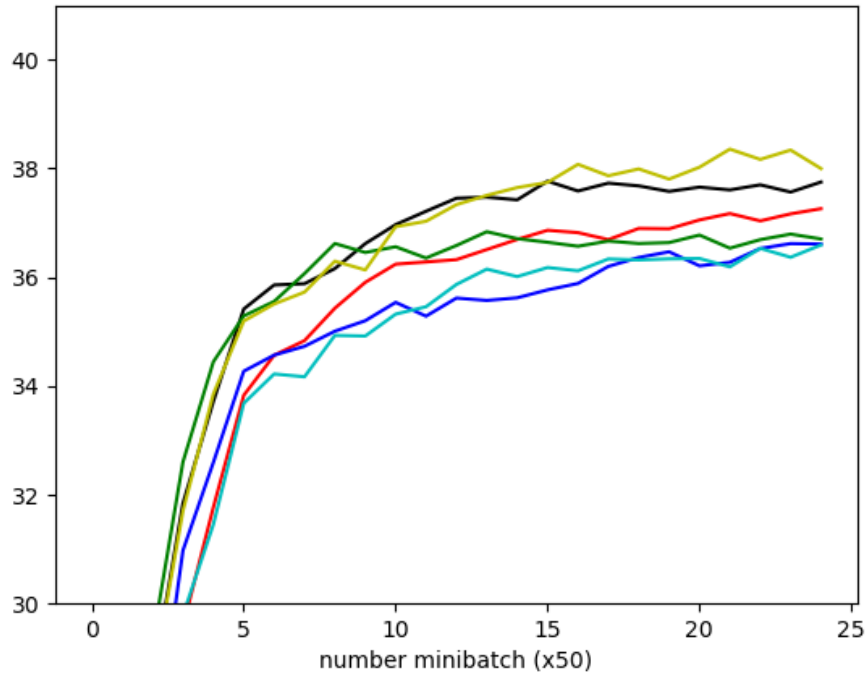


(f) Online VB

Tập dữ liệu Tweets, K = 100, sử dụng độ đo NPMI



(g) Online Gibbs sampling

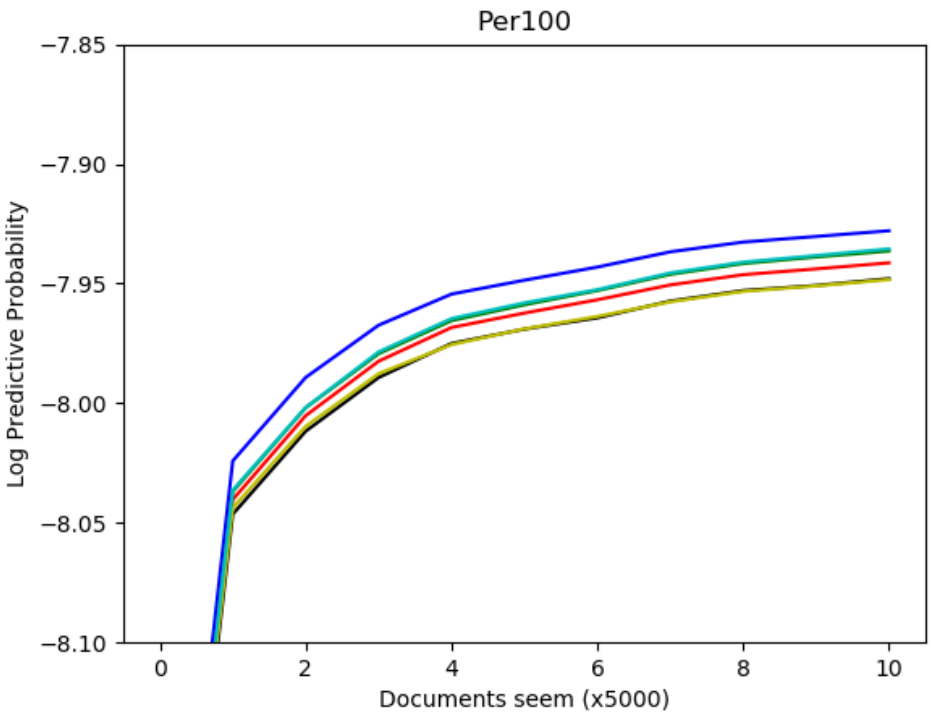


(h) Online VB

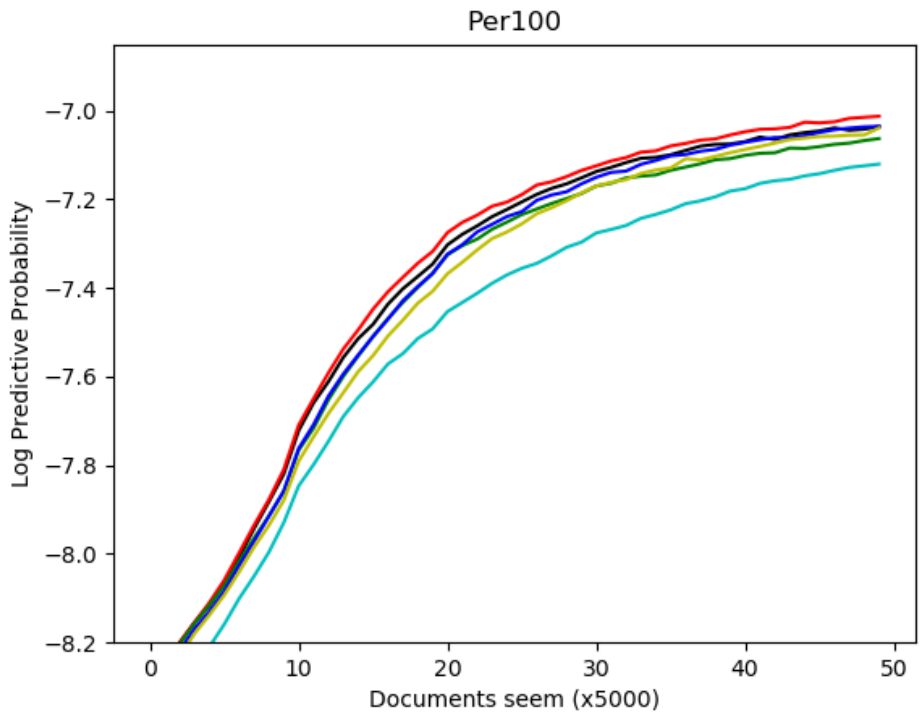
Thời gian chạy với bộ dữ liệu Tweets

drop rate	K	0	0.1	0.3	0.5	0.7	0.9
Gibbs	50	249935	231575	186822	80132	56739	24534
VB	50	3592	3463	2922	2166	1249	375
Gibbs	100	404573	357891	288019	213525	134510	46840
VB	100	3224	3050	2503	1870	1189	880
Gibbs	150	460189	429685	350120	297817	296984	114108
VB	150	3866	3696	3052	2429	1536	781
Gibbs	200	539904	538529	456806	387955	275327	171603
VB	200	4925	4594	3313	2622	1944	1322

Tập dữ liệu Yahoo, K =100, độ đo perplexity

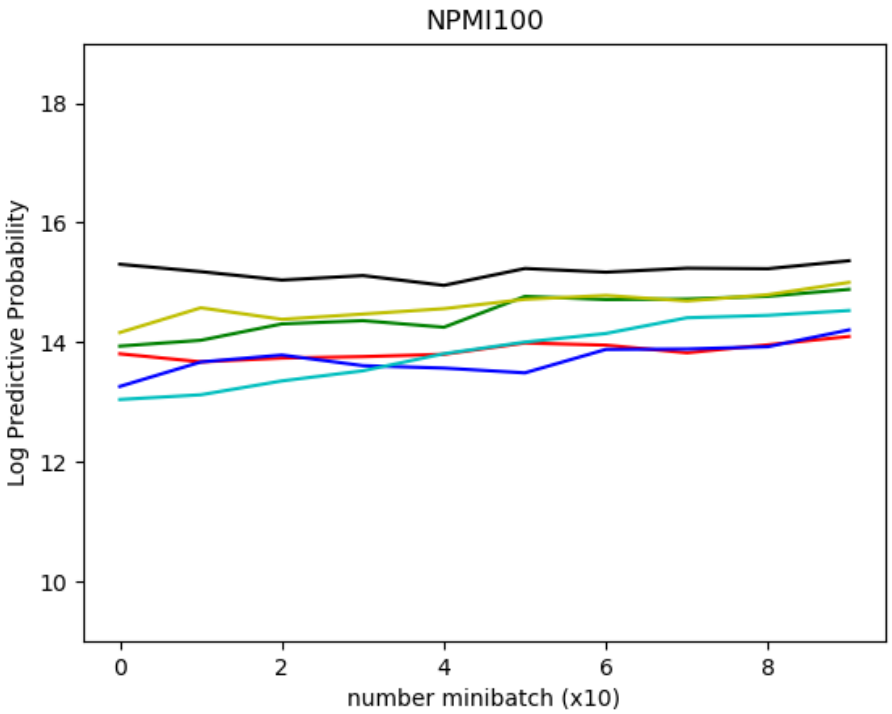


(i) Online Gibbs sampling

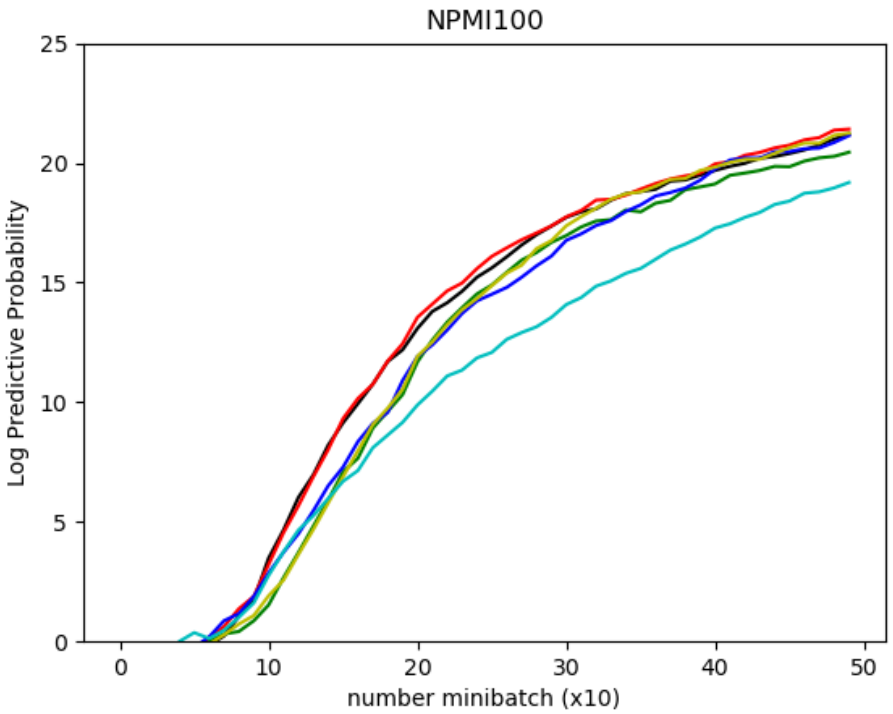


(j) Online VB

Tập dữ liệu Yahoo, K = 100, sử dụng độ đo NPMI



(k) Online Gibbs sampling

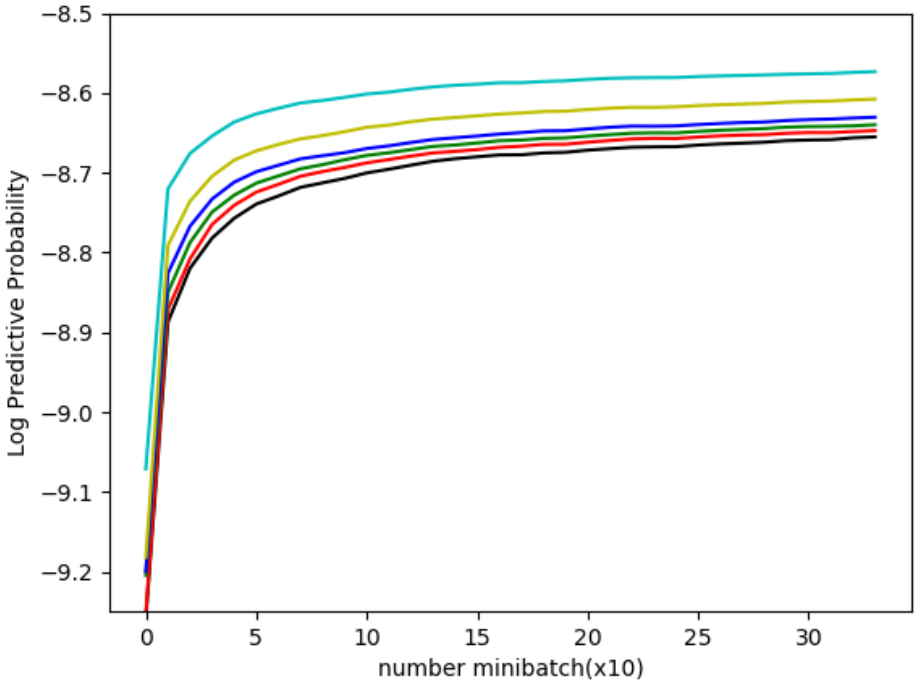


(l) Online VB

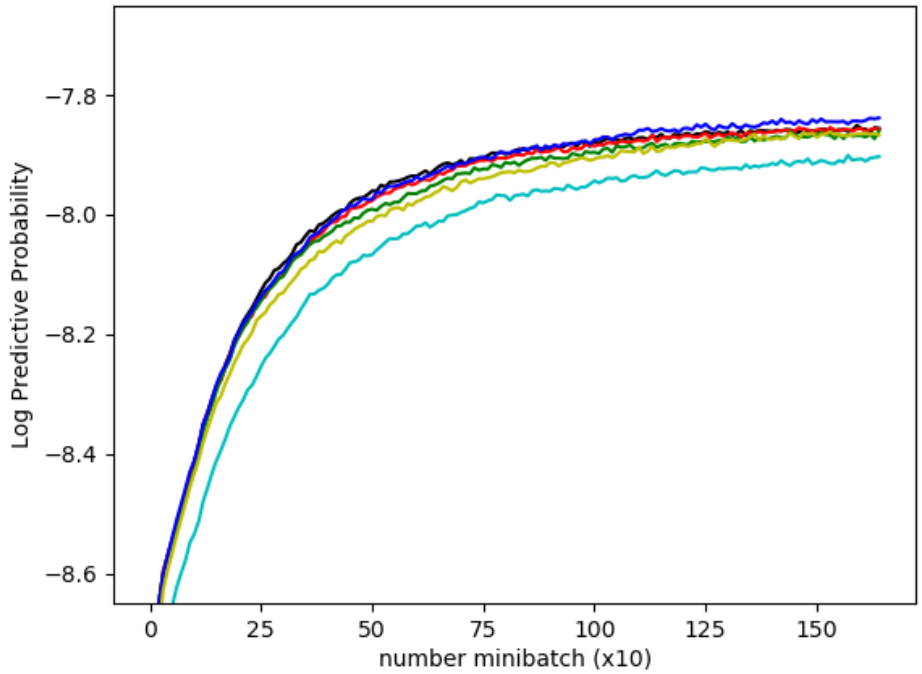
Thời gian chạy với bộ dữ liệu Yahoo

drop rate	K	0	0.1	0.3	0.5	0.7	0.9
Gibbs	50	24259	22605	18919	14154	9594	3856
VB	50	3592	3463	2922	2166	1249	375
Gibbs	100	32042	27544	21763	16729	19533	7565
VB	100	448	418	372	328	274	223
Gibbs	150	73385	68762	58344	44619	30748	12387
VB	150	514	490	453	398	337	265
Gibbs	200	64068	51862	46133	6406	41653	8894
VB	200	580	559	485	434	373	306

Tập dữ liệu NYT, K = 100, sử dụng độ đo perplexity

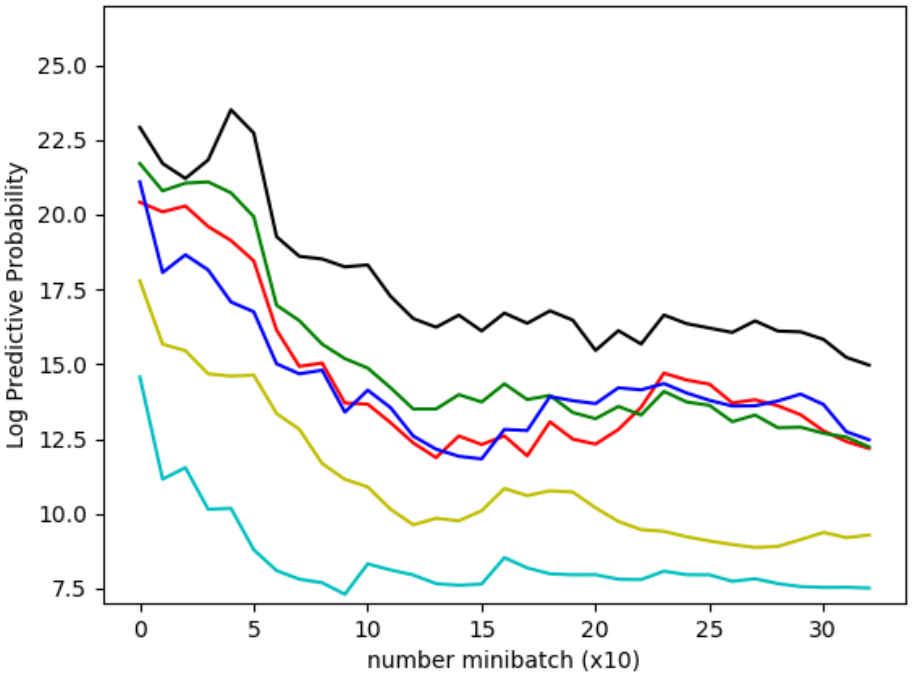


(m) Online Gibbs sampling

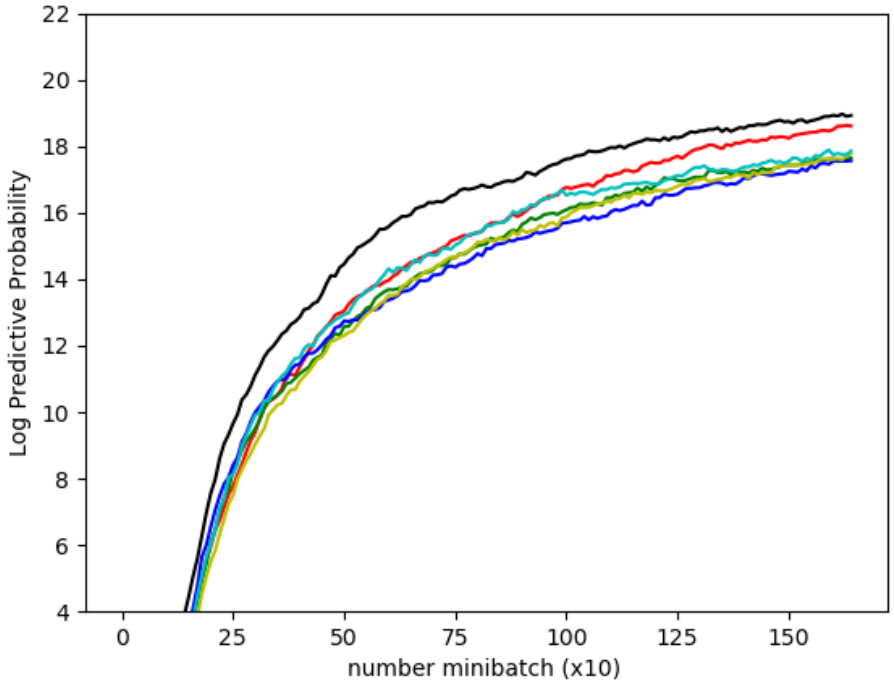


(n) Online VB

Tập dữ liệu NYT, K = 100, sử dụng độ đo NPMI



(o) Online Gibbs sampling



(p) Online VB

Thời gian chạy với bộ dữ liệu NYT

drop rate	K	0	0.1	0.3	0.5	0.7	0.9
Gibbs	50	81029	75909	63424	49757	33146	11787
VB	50	3592	3463	2922	2166	1249	375
Gibbs	100	90219	85610	71573	56306	39515	15900
VB	100	3224	3050	2503	1870	1189	880
Gibbs	150	130083	123790	124131	107083	89761	38267
VB	150	3866	3696	3052	2429	1536	781
Gibbs	200	193110	180111	184658	151754	47640	33396
VB	200	4925	4594	3313	2622	1944	1322

Đánh giá

- Phương pháp Gibbs sampling
  - Thời gian chạy rất chậm.
  - Sử dụng phương pháp học ngẫu nhiên cho kết quả xấp xỉ nhau nhưng thời gian cải thiện rất đáng kể.
- Phương pháp VB cho thời học rất nhanh, và kết quả cao hơn so với phương pháp học Gibbs sampling

Nội dung

Kết luận

- Hai phương pháp học mới:
  - Phương pháp học VB
    => Cho kết quả cao hơn và thời gian nhanh hơn rất nhiều so với phương pháp Gibbs sampling
  - Phương pháp học ngẫu nhiên
    => Cải thiện thời gian học rất đáng kể